In **Proc. First Int. Workshop on Camera-Based Document Analysis and Recognition**, Seoul, Korea, 2005.

Experiments in Video-Based Whiteboard Reading

Gernot A. Fink Markus Wienecke[1] Gerhard Sagerer
Bielefeld University, Faculty of Technology
33594 Bielefeld, Germany
gernot@techfak.uni-bielefeld.de

Abstract

With the increasing computational support for collaborative work-environments electronically enhanced whiteboards have been developed to serve as automatic meeting assistants. The most flexible of these systems use cameras to observe the whiteboard, and, therefore, do not require the use of special pens or erasers. However, currently these systems are only capable to interpret some special graphical symbols and can not produce transcripts of the documents written on them. As a major advancement beyond the state-of-the-art we propose a system for automatic video-based reading of unconstrained handwritten text from a whiteboard. Text lines are extracted from the captured image sequence using an incremental processing strategy. The recognition results are then obtained from the text-line images by off-line techniques and a segmentation-free statistical recognizer. We will present results on a writer independent unconstrained handwriting recognition task showing that handwriting recognition can successfully be applied to automatically reading texts from whiteboards.

1. Introduction

Whiteboards are very popular tools not only for presentations and educational purposes but also in meeting rooms for the exchange of ideas during group discussions, for project planning, system design, etc.

In order to make use of whiteboards as user interfaces for human computer interaction in such collaborative working environments, systems based on electronic whiteboards have been developed. Similar to digitizing tablets these systems employ electronic pens and erasers allowing their positions in the plane to be sensed and tracked during the writing process. This data can then be used to construct an electronic version of the document-image on the whiteboard. Additionally, the pen trajectory can be interpreted by an on-line recognition module to automatically recognize what was written on the board.

However, electronic whiteboards exhibit some disadvantages. As special pens and erasers are necessary, the natural interaction is restricted. Therefore, a promising alternative is to retain ordinary whiteboards and pens, and to observe the writing process using a video camera.

In order to cover a large area of the whiteboard, the preferable position of the camera is several meters in front of the board, either mounted to the ceiling or fixed on a tripod. However, in such a setup the writer will usually stand in front of the board while writing. Therefore, the pen and portions of text are frequently occluded by the user. In order to circumvent this drawback, a kind of activity analysis could be employed to decide whether the captured image is suitable for further processing. An alternative method is to extract only the visible portions of the handwritten text and to incrementally integrate the partial transcriptions into the overall recognition result.

In our previous work we proposed a system for automatic video-based whiteboard reading [14]. In contrast to the approaches presented in [9, 10], which only permit the recognition of a limited set of symbols, our system is designed for recognizing unconstrained handwritten text. As the pen is rarely visible in the image and thus on-line recognition based on the pen trajectory is not feasible, an incremental off-line recognition approach is applied. In this paper we will present results of a thorough evaluation of our system on a writer independent unconstrained handwriting recognition task that clearly demonstrate that handwriting recognition can successfully be applied to automatically reading texts from whiteboards.

In the following section we will briefly review some relevant related work. In section 3 we will give an overview of the architecture of the proposed whiteboard reading system. The techniques applied for statistical modeling and recognition of unconstrained handwritten texts will be described in

[1] Markus Wienecke now is with Siemens AG, Logistics and Assembly Systems, Postal Automation Division, Constance, Germany.

[2] An extended version of this paper will appear in *Int. Journal on Document Analysis and Recognition* (IJDAR).

detail in section 4. Finally, the results of our extensive evaluation experiments will be presented in section 5.

2. Related Work

Electronically enhanced whiteboards that do not make use of specialized hardware for pen tracking, observe the board with cameras. In these systems pen movements or relevant image regions have to be extracted from the captured image sequences.

One approach is to use a special marker for writing that has a distinctive color. By tracking that pen a temporal trajectory is obtained that can be recognized using on-line methods. In [1] such a system is described, which allows the user to control a computer with simple gestures produced by a special marker pen. A video-based system which is capable to track an ordinary pen in image sequences was proposed by Munich & Perona [8]. In [4] the trajectories generated by this approach were used for on-line handwriting recognition.

On-line type systems as these can be successfully employed in scenarios where the pen is always visible in the image. However, they can hardly be applied for whiteboard reading where the pen is very often occluded by the writer.

Therefore, a contrary approach for video-based whiteboard reading is to extract and analyze the relevant image regions after the writing process has finished. For example, the video-based *BrightBoard* system described in [10] continuously observes the whiteboard and grabs a suitable image when the motion of the writer stops. This image is analyzed in order to find and recognize graphical marks that correspond to control commands. A similar camera-based whiteboard scanner is the so-called *ZombieBoard* system proposed in [9], which applies a mosaicing algorithm to enable high-resolution imaging. The system monitors activity in front of the board and detects the drawing of graphical marks indicating commands and associated parameters.

As the system presented in this paper is not restricted to a small set of commands but is designed for recognizing unconstrained handwritten text the approach is also closely related to the task of off-line handwriting recognition (cf. e.g. [11]). In contrast to isolated word recognition the task of recognizing unconstrained handwritten texts using a large or even unlimited vocabulary is much more difficult. This is mainly caused by the absence of context knowledge and word segmentation information. Most current systems, therefore, rely on segmentation-free methods in order to avoid errors introduced by segmenting the text into words or even characters at an early stage. Especially, Hidden-Markov Models (HMMs) were successfully applied and gained growing interest in the research community. Advanced systems for writer-independent unconstrained text recognition are described, e.g., in [5, 6, 12] or [13].

3. System Overview

Our system for automatic whiteboard reading applies an incremental processing strategy. The writing process is continuously observed and the recognition process is activated as soon a handwritten text region is visible in the image. Thus, the text regions are transcribed in their order of appearance and integrated into the overall recognition result.

The writing process is observed with a video camera positioned approximately 3 m from the whiteboard. After grabbing an image that shows an area of approximately 70×50 cm all text regions currently visible are extracted. In order to avoid recognizing the same text region multiple times in the image sequence, we employ a region memory containing all the different text regions extracted so far.

If a new, not yet memorized text region is found, several pre-processing steps are applied to compensate for the highly varying background intensity and to normalize the handwriting. First, the region image is binarized using an adaptive threshold that depends on the intensity distribution in a local neighborhood. Then the the vertical position, skew, and slant of each text region are corrected locally. Especially for the baseline estimation a local procedure is absolutely crucial as in texts written on a whiteboard frequently baseline drifts can be observed (see figure 1). In order to make the subsequent feature extraction process more robust, finally, the size of the handwriting is normalized. This is achieved by rescaling the line images such that the average distance between local extrema of the text contour matches a predefined distance.

From the pre-processed line images a set of 9 geometric features is extracted in a sliding window technique similar to the approach described in [6]. For considering a wider context, we additionally compute an approximate horizontal derivative for each component of the feature vector, so that an 18 dimensional feature vector is obtained. The details of the text extraction, preprocessing, and feature extraction methods applied can be found in [14].

4. Statistical Modeling & Recognition

A successful statistical recognition system for handwriting or spoken language consists of two modeling components, one that describes the realization of individual segments, e.g. words or characters, and another that describes the restrictions on the expected segment sequences. The first component is usually realized by HMMs that model the probability density $p(\boldsymbol{x}|\boldsymbol{w})$ of observing a certain sequence of feature vectors $\boldsymbol{x}$ given a sequence of words or characters $\boldsymbol{w}$. The restriction of these sequences to plausible ones is achieved by defining a probability distribution $P(\boldsymbol{w})$ for all possible sequences, which can be realized by a Markov-chain or n-gram model. The goal of the recogni-

	Source	Type	Categories	Documents	Writers	Lines	Words	Characters
Training	IAM-DB	scanned document	A – D	492	>200	4222	36582	189852
		text prompt	A – D	492	–	–	37273	–
Cross-validation	IAM-DB	scanned document	E – F	129	≈50	1081	9612	49002
Test	whiteboard	video document	F01	20	10	173	1171	6171

Table 1. Corpora of handwritten & text data: word counts include punctuation and word fragments resulting from hyphenation; character counts include approximately 20% of white space.

tion process is then to find the word or character sequence $\hat{\boldsymbol{w}}$ that maximizes the probability of the combined statistical model given the observed data $\boldsymbol{x}$ according to:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, p(\boldsymbol{x}|\boldsymbol{w})P(\boldsymbol{w})$$

In analogy to the terminology used in spoken language processing the HMM $p(\boldsymbol{x}|\boldsymbol{w})$ could be termed the *writing model* and the n-gram model $P(\boldsymbol{w})$ is equivalent to the so-called *language model*.

4.1. Corpora

For the design of statistical recognition systems the availability of a sufficiently large database of training samples is an important prerequisite. Ideally, for a video-based system it would be desirable to obtain a large amount of image data recorded while observing a subject writing on the whiteboard. However, recording and labeling of such video data requires a substantial manual effort. Therefore, we decided to use the IAM-database of scanned documents [7] for training and cross-validation. The database provides a large amount of handwritten text documents that were produced by several hundred subjects. The documents are divided into categories according to the different topics covered.

Unfortunately, the IAM-database does not contain writer IDs for the handwritten samples. However, writers never provided samples for different categories. Therefore, we defined the training data to comprise categories A to D and the cross-validation data categories E & F. This partitioning corresponds to the training and test sets used in [13] and ensures all experiments to be truly writer independent.

The test data was collected in our lab by recording image sequences of texts written on a whiteboard. In order to be able to compare the performance of the video-based system with our off-line recognizer [13], we asked ten subjects to write portions from the off-line cross-validation texts on the whiteboard, namely from category F01. No constraints with respect to the writing style were given. In contrast to the training patterns resulting from scanned forms, where rulers on a second sheet put below were used to align the baseline horizontally, the video-based data often shows baseline drifts and variations of the corpus height.

A summary of the relevant characteristics of the corpora used is given in table 1. Figure 1 shows examples of a scanned document used for training and the final version of a video document from the test data. Additionally, the results of the incremental text detection are shown, which, for the example given, produces the lines in a different order than found in the final document.

4.2. Writing Model

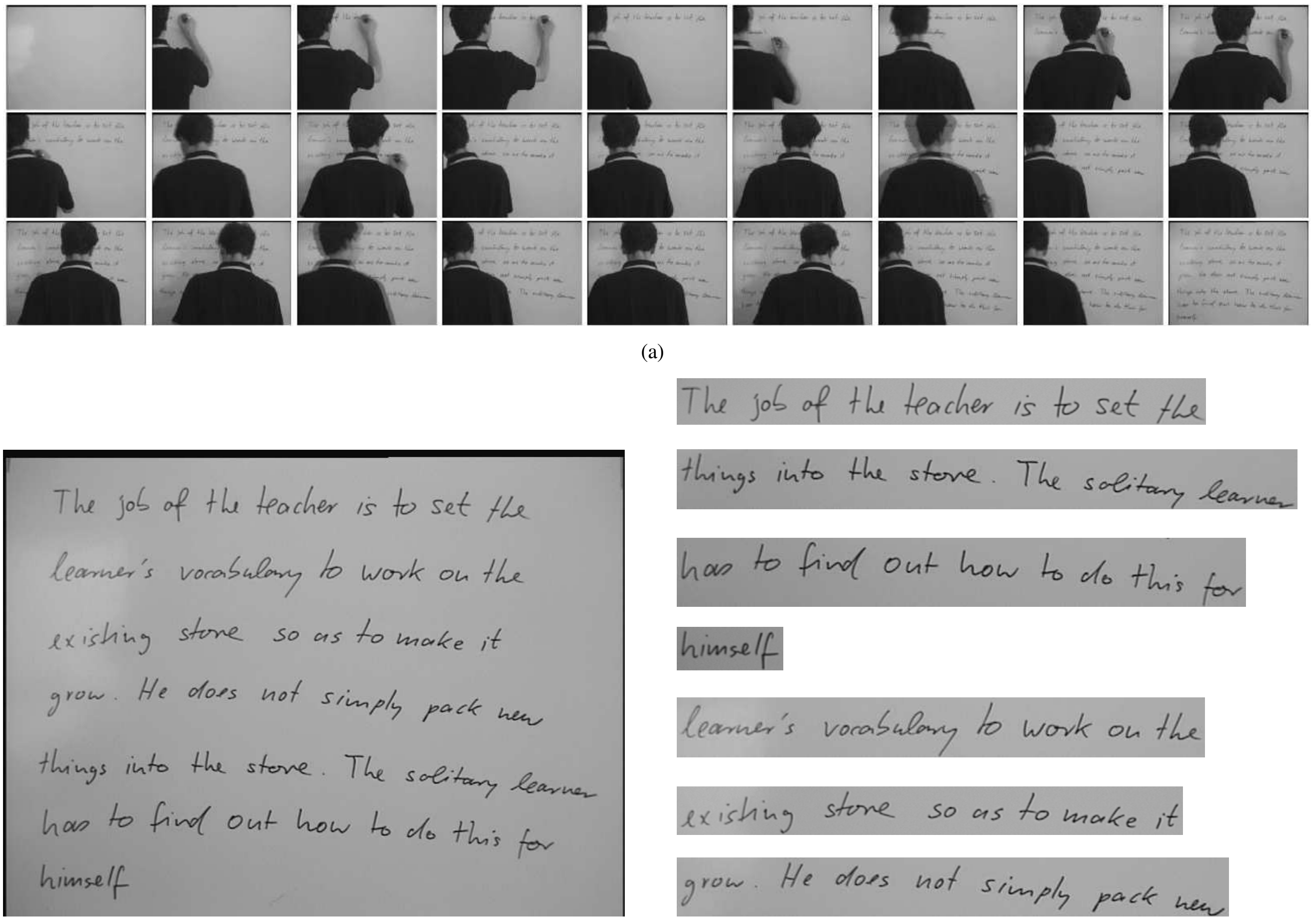
The configuration and parameter estimation for the HMMs defining the writing model as well as for the language models used is carried out in the framework of the ESMERALDA development environment [3].

As general setup we use semi-continuous HMMs with a shared codebook of approximately 2000 Gaussian mixtures with diagonal covariance matrices. A total of 75 HMMs are created for modeling 52 letters, ten numbers, twelve punctuation marks and brackets, and white space. The latter consists of three variants accounting for different lengths in blank space between words or characters. All these models use the *Bakis*-type topology, i.e. they are basically linear models which in addition to loops and forward state-transitions permit the skipping of states in the sequence. Thus, the models can cope with a wider range of lengths in the character patterns described.

The shared codebook is initialized in un-supervised mode by applying the k-means algorithm to the training data. Then the initial HMM parameters can be determined on labeled initialization data. Afterwards, we apply several iterations of the Baum-Welch parameter re-estimation to the models. From the context-independent character model set thus obtained, models for arbitrary words of some given lexicon can be constructed easily by concatenating the appropriate character models.

4.3. Language Model

For estimating character-based language models the transcriptions of the training data and for word-based mod-



(a)

The job of the teacher is to set the
learner's vocabulary to work on the
existing store so as to make it
grow. He does not simply pack new
things into the store. The solitary learner
has to find out how to do this for
himself

(b)

The job of the teacher is to set the

things into the store. The solitary learner

has to find out how to do this for

himself

learner's vocabulary to work on the

existing store so as to make it

grow. He does not simply pack new

(c)

Figure 1. Examples of whiteboard data collected: (a) image sequence captured (summarized), (b) resulting final whiteboard document, and (c) text lines extracted during observation of the writing process (in order of extraction).

els the original text prompts were used. The raw n-gram probability distributions were smoothed by applying absolute discounting and backing-off (cf. e.g. [2]).

A major limitation for the performance of a word-based language model in our configuration of training and test data arises from the fact that the texts belong to different categories covering widely differing topics. From the total of 2534 word forms appearing in the text prompts of the cross-validation data (categories E & F) more than 48% never appeared in the training texts (categories A – D). Additionally, writers sometimes used varying hyphenation which introduces unseen word fragments. In the whiteboard data 316 different word forms are used, more than 26% of which are not covered by the training set. Therefore, we decided to include in addition to the lexicon of the training data all those word forms in the overall recognition lexicon that are necessary to describe the text prompts from which cross-validation and test set were generated. From this word list a small number of entries was eliminated, which contained characters not present in the training material. The resulting recognition lexicon consists of 7485 entries including punctuation and word fragments resulting from hyphenation. The percentage of out-of-vocabulary words for both cross-validation and test data is approximately 0.5%.

5. Results & Discussion

In order to evaluate the proposed methods for video-based whiteboard reading we carried out several experiments on the test set described in section 4.1. Whenever possible the results obtained are compared to those achieved by an off-line recognition system on the cross-validation data. A comparison of those figures with results on data from the IAM-database reported in the literature [5, 6, 12] also clearly shows the excellent quality achieved in modeling and decoding. Though restricting the possible occurrence of upper-case characters to word-initial positions Kavallieratou and colleagues only achieved character error rates slightly below 30% on the IAM-database [5]. With a 7k vocabulary and a bi-gram language model Marti *et. al* achieve a word error rate of approximately 40% [6]. Vinciarelli *et. al* report error rates between 57% and 55% for different language models ranging from uni- to tri-gram when using a 10k lexicon [12].

5.1. Text Detection

The precondition for whiteboard reading is to robustly detect the image regions of the handwriting. Therefore, we first investigated the effectiveness of the method for text detection. Using the 20 image sequences for testing consisting of 152 handwritten lines of text, it turned out that a total of 188 image regions have been detected. 173 of these regions are correctly detected text regions. In only 15 cases errors occurred due to noise or line segmentation errors caused by touching or heavily overlapping lines. The discrepancy of the total number of originally written lines (152) and the overall number of correctly detected text regions (173) is caused by the incremental processing strategy. Thus, we observed that in 21 cases portions of text lines have been detected repeatedly. Additionally, we investigated whether the sequence of detected regions corresponds to the chronological order in which the text lines were written on the board. From the overall number of 173 text regions the chronological order was not correct in 9 cases (see e.g. figure 1).

5.2. Lexicon-based Recognition

For lexicon-based recognition of whiteboard texts we used a lexicon containing 7485 word forms (see section 4.3). The results achieved are summarized in table 2. Without the use of any restrictions on the possible word sequences we obtain a word error rate of 47.8%. Clearly, such a figure would not be acceptable for an automatic transcription system. However, with some limited knowledge about the expected texts represented as a bi-gram language model this figure could be improved to 28.9%. This corresponds to a reduction of the error rate of approximately 40%. Due to the widely differing lexicons of training and test data the bi-gram model has a very high perplexity on both test and cross-validation set. For a well trained language model that could be estimated on text data *matching* the topics of the final application a substantially lower perplexity can be expected[1]. Therefore, word-based recognition results on white-board data could easily be improved further for better matching training and test conditions.

	% WER / perplexity	
	none	2-gram
Cross-validation	43.9 / (7485)	28.3 / 757
Test (whiteboard)	47.8 / (7485)	28.9 / 645

Table 2. Word error rates (WER) achieved for a 7485-word lexicon with and without using a bi-gram language model.

5.3. Lexicon-free Recognition

Ultimately, any handwriting recognition system should be able to recognize text independently from a predefined list of possible words. For such lexicon-free recognition at least some expectation on the possible sequence of characters is required.

Therefore, we estimated character-based language models with n-gram lengths ranging from two to five (see section 4.3). These models were then used in conjunction with the context-independent character HMMs during the recognition process. The results obtained are shown in table 3. Without the restriction of a language model a character error rate of 31.0% is obtained, i.e. roughly every third character – including white space – is misrecognized. However, when using the statistical restrictions on possible character sequences as represented by the character based language models this figure can be improved significantly. With a 5-gram model a character error rate as low as 19.0% can be achieved on the whiteboard data.

Though the mismatch of lexicons between training and test data is a severe limitation for word-based recognition it has an advantage for the judgment of the lexicon-free results. In principle long-span n-gram models could learn the training lexicon and, therefore, results obtained with such a model might not be truly lexicon-free. In our configuration, however, learning of the word forms found in the training texts has very limited effect on the cross-validation and test data (see also section 4.3). Therefore, the low character error rates achieved impressively demonstrate the capability of the n-gram models to capture more general characteristics of the character sequences.

5.4. Video vs. Off-line Recognition

The comparison of the recognition results obtained on the whiteboard data and on the scanned documents used for cross-validation clearly shows better performance on the

[1] Despite a lower perplexity ($\approx$ 400) when using a 10k lexicon and a bi-gram language model in [12] only a surprisingly high error rate of almost 60% is achieved on data from the IAM-database.

	% CER / perplexity				
	none	2	3	4	5
Cross-validation	29.2 / (75)	22.1 / 12.7	18.3 / 9.3	16.1 7.7	15.6 / 7.3
Test (whiteboard)	31.0 / (75)	25.9 / 12.0	22.0 / 8.5	20.1 / 6.9	19.0 / 6.5

Table 3. Character error rates (CER) achieved with different n-gram language models.

latter ones. However, the difference in recognition quality is relatively small when considering the widely different nature of the documents used. This evidence makes it obvious that the methods used for text-detection, preprocessing and feature extraction are capable of compensating for the majority of distortion effects found in the video data.

6. Conclusion

We presented a system for automatic whiteboard reading based on visual input. It is characterized by an incremental processing strategy, i.e. the text lines are extracted as soon as they are visible in the image. The pre-processing and feature extraction methods applied generate a data representation which is to a certain extent robust against variations concerning the writing style and the reduced quality of the video-based data. Evaluation results on a writer independent task were presented for both lexicon-based and lexicon-free recognition of unconstrained handwriting. When using a 7.5k lexicon and a bi-gram model a word error rate of only 28.9% could be achieved. Without an explicit lexicon and the use of only a character 5-gram model a character error rate as low as 19.0% was reached. These results clearly demonstrate the effectiveness of the proposed methods for text detection, preprocessing, feature extraction, and statistical modeling and recognition and their successful combination in a complete system for automatic video-based whiteboard reading.

Acknowledgments

This work was supported by the German Research Foundation (DFG) within project **Fi799/1**.

Additionally, we would like to thank the Institute of Informatics and Applied Mathematics, University of Bern, namely Horst Bunke and Urs-Viktor Marti, who allowed us to use the IAM database of handwritten forms [7] for our recognition experiments.

References

[1] M. J. Black and A. D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In H. Burkhardt and B. Neumann, editors, *European Conf. on Computer Vision*, pages 909–924, Freiburg, Germany, 1998.

[2] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13:359–394, 1999.

[3] G. A. Fink. Developing HMM-based recognizers with ESMERALDA. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Text, Speech and Dialogue*, volume 1692 of *Lecture Notes in Artificial Intelligence*, pages 229–234. Springer, Berlin Heidelberg, 1999.

[4] G. A. Fink, M. Wienecke, and G. Sagerer. Video-based online handwriting recognition. In *Proc. Int. Conf. on Document Analysis and Recognition*, pages 226–230, 2001.

[5] A. Kavallieratou, N. Fakotakis, and G. Kokkinakis. An unconstrained handrwiting recognition system. *Int. Journal on Document Analysis and Recognition*, 4:226–242, 2005.

[6] U.-V. Marti and H. Bunke. Handwritten sentence recognition. In *Proc. Int. Conf. on Pattern Recognition*, volume 3, pages 467–470, Barcelona, 2000.

[7] U.-V. Marti and H. Bunke. The IAM-database: An english sentence database for offline handwriting recognition. *Int. Journal on Document Analysis and Recognition*, 5(1):39–46, 2002.

[8] M. E. Munich and P. Perona. Visual input for pen-based computers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(3):313–328, 2002.

[9] E. Saund. Bringing the marks on a whiteboard to electronic life. In *Proc. 2nd Int. Workshop on Cooperative Buildings, CoBuild'99*, pages 69–78, Pittsburgh, 1999. Springer.

[10] Q. Stafford-Fraser and P. Robinson. Brightboard: A video-augmented environment. In *Proc. Conf. on Human Factors and Computing Systems*, pages 134–141, Vancouver, BC, Canada, 1996.

[11] T. Steinherz, E. Rivlin, and N. Intrator. Offline cursive script word recognition – A survey. *Int. Journal on Document Analysis and Recognition*, 2(2):90–110, 1999.

[12] A. Vinciarelli, S. Bengio, and H. Bunke. Offline recognition of unconstrainded handwritten texts using HMMs and statistical language models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(6):709–720, 2004.

[13] M. Wienecke, G. A. Fink, and G. Sagerer. Experiments in unconstrained offline handwritten text recognition. In *Proc. 8th Int. Workshop on Frontiers in Handwriting Recognition*, Niagara on the Lake, Canada, August 2002.

[14] M. Wienecke, G. A. Fink, and G. Sagerer. Towards automatic video-based whiteboard reading. In *Proc. Int. Conf. on Document Analysis and Recognition*, pages 87–91, Edinburgh, 2003.